# Annotating Underquantification

**Aurelie Herbelot**
University of Cambridge
Cambridge, United Kingdom
ah433@cam.ac.uk

**Ann Copestake**
University of Cambridge
Cambridge, United Kingdom
aac10@cam.ac.uk

## Abstract

Many noun phrases in text are ambiguously quantified: syntax doesn't explicitly tell us whether they refer to a single entity or to several, and what portion of the set denoted by the Nbar actually takes part in the event expressed by the verb. We describe this ambiguity phenomenon in terms of underspecification, or rather **underquantification**. We attempt to validate the underquantification hypothesis by producing and testing an annotation scheme for quantification resolution, the aim of which is to associate a single quantifier with each noun phrase in our corpus.

## 1 Quantification resolution

We are concerned with **ambiguously quantified** noun phrases (NPs) and their interpretation, as illustrated by the following examples:

1. Cats are mammals = *All* cats...
2. Cats have four legs = *Most* cats...
3. Cats were sleeping by the fire = *Some* cats...
4. The beans spilt out of the bag = *Most/All of the* beans...
5. Water was dripping through the ceiling = *Some* water...

We are interested in **quantification resolution**, that is, the process of giving an ambiguously quantified NP a formalisation which expresses a *unique* set relation appropriate to the semantics of the utterance. For instance, we wish to arrive at:

6. All cats are mammals.

   $|\phi \cap \psi| = |\phi|$ where $\phi$ is the set of all cats and $\psi$ the set of all mammals.

Resolving the quantification value of NPs is important for many NLP tasks. Let us imagine an information extraction system having retrieved the triples 'cat – is – mammal' and 'cat – chase – mouse' for inclusion in a factual database about felines. The problem with those representation-poor triples is that they do not contain the necessary information about quantification to answer such questions as 'Are all cats mammals?' or 'Do all cats chase mice?' Or if they attempt to answer those queries, they give the same answer to both. Ideally, we would like to annotate such triples with quantifiers which have a direct mapping to probability adverbs:

7. *All* cats are mammals AND Tom is a cat → Tom is *definitely* a mammal.
8. *Some* cats chase mice AND Tom is a cat → Tom *possibly* chases mice.

Adequate quantification is also necessary for inference based on word-level entailment: an existentially quantified NP can be replaced by a suitable hypernym but this is not possible in non-existential cases: *(Some) cats are in my garden* entails *(Some) animals are in my garden* but *(All) cats are mammals* doesn't imply that *(All) animals are mammals*.

In Herbelot (to appear), we provide a formal semantics for ambiguously quantified NPs, which relies on the idea that those NPs exhibit an underspecified quantifier, i.e. that for each NP in a corpus, a set relation can be agreed upon. Our formalisation includes a placeholder for the quantifier's set relation. In line with inference requirements, we assume a three-fold partitioning of the quantificational space, corresponding to the natural language quantifiers *some*, *most* and *all* (in addition to *one*, for the description of singular, unique entities). The corresponding set relations are:

9. $some(\phi, \psi)$ is true iff $0 < |\phi \cap \psi| < |\phi - \psi|$
10. $most(\phi, \psi)$ is true iff $|\phi - \psi| \leq |\phi \cap \psi| < |\phi|$
11. $all(\phi, \psi)$ is true iff $|\phi \cap \psi| = |\phi|$

This paper is an attempt to show that our formalisation lends itself to evaluation by human annotation. The labels produced will also serve as training and test sets for an automatic quantification resolution system.



*Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010*, pages 73–81,
Uppsala, Sweden, 15-16 July 2010. 

## 2 Under(specified) quantification

Before we present our annotation scheme, we will spell out the essential idea behind what we call **underquantification**.

The phenomenon of ambiguous quantification overlaps with **genericity** (see Krifka et al, 1995, for an introduction to genericity). Generic NPs are frequently expressed syntactically as bare plurals, although they occur in definite and indefinite singulars too, as well as bare singulars. There are many views on the semantics of generics (e.g. Carlson, 1995; Pelletier and Asher, 1997; Heyer, 1990; Leslie, 2008) but one of them is that they quantify (Cohen, 1996), although, puzzlingly enough, not always with the same quantifier:

12. Frenchmen eat horsemeat = *Some/Relatively-many* Frenchmen... (For the *relatively many* reading, see Cohen, 2001.)
13. Cars have four wheels = *Most* cars...
14. Typhoons arise in this part of the Pacific = *Some* typhoons... OR *Most/All* typhoons...

This behaviour has so far prevented linguists from agreeing on a single formalisation for all generics. The only accepted assumption is that an operator $GEN$ exists, which acts as a silent quantifier over the restrictor (subject) and matrix (verbal predicate) of the generic statement. The formal properties of $GEN$ are however subject to debate: in particular, it is not clear which natural language quantifier it would map onto (some view it as *most*, but this approach requires some complex domain restriction to deal with sentences such as 12).

In this paper, we take a different approach which sidesteps some of the intractable problems associated with the literature on generics and which also extends to definite plurals. Instead of talking of ambiguous quantification, we will talk of **underspecified quantification**, or **underquantification**. By this, we mean that the bare plural, rather than exhibiting a silent, $GEN$ quantifier, simply features a placeholder in the logical form which must be filled with the appropriate quantifier (e.g., $uq(x, \text{cat}'(x), \text{sleep}'(x))$, where $uq$ is the placeholder quantifier). This account caters for the facts that so-called generics can so easily be quantified via traditional quantifiers, that $GEN$ is silent in all known languages, and it explains also why it is the bare form which has the highest productivity, and can refer to a range of quantified sets, from existentials to universals. Using the underquantification hypothesis, we can paraphrase any generic of the form 'X does Y' as 'there is a set of things X, *a certain number of which* do Y' (note the partitive construction). Such a paraphrase allows us to also resolve ambiguously quantified definite plurals, which have traditionally been associated with universals, outside of the genericity phenomenon (e.g. Lyons, 1999).

Because of space constraints, we will not give our formalisation for underquantification in this paper (see Herbelot,to appear, for details). It involves a representation of the partitive construct exemplified above and requires knowledge of the distributive or collective status of the verbal predicate. We also argue that if generics can always be quantified, their semantics may involve more than quantification. So we claim that in certain cases, a double formalisation of the NP as a quantified entity and a kind is desirable. We understand kinds in the way proposed by Chierchia (1998), that is as the plurality of all instances denoted by a given word in the world under consideration. Under the kind reading, we can interpret 12 as meaning *Collectively, the group of all Frenchmen has the property of eating horsemeat*.

## 3 Motivation

### 3.1 Linguistic motivation

It is usual to talk of 'annotation' generically, to cover any process that involves humans using a set of guidelines to mark some specific linguistic phenomenon in some given text. However, we would argue that, when considering the aims of an annotation task and its relation to the existing linguistic literature, it becomes possible to distinguish between various types of annotation. Further, we will show that our own effort situates itself in a little studied relation to formal semantics.

The most basic type of annotation is the one where computational linguists mark large amounts of textual data with well-known and well-understood labels. The production of tree banks like the Penn Treebank (Marcus et al, 1993) makes use of undisputed linguistic categories such as parts of speech. The aim is to make the computer learn and use irrefutable bits of linguistics. (Note that, despite agreement, the representation of those categories may differ: see for example the range of available parts of speech tag sets.) This type of task mostly involves basic syntactic knowledge, but can be taken to areas of syntax and seman-

tics where the studied phenomena have a (somewhat) clear, agreed upon definition (Kingsbury et al, 2002). We must clarify that in those cases, the choice of a formalism may already imply a certain theoretical position – leading to potential incompatibilities between formalisms. However, the categories for such annotation are themselves fixed: there is a generally agreed broad understanding of concepts such as noun phrases and coordination.

Another type of annotation concerns tasks where the linguistic categories at play are not fixed. One example is discourse annotation according to rhetorical function (Teufel et al, 2006) where humans are asked to differentiate between several discursive categories such as 'contrast' or 'weakness'. In such a task, the computational linguist develops a theory where different states or values are associated with various phenomena. In order to show that the world functions according to the model presented, experimentation is required. This usually takes the form of an annotation task where several human subjects are required to mark pieces of text following guidelines inferred from the model. The intuition behind the annotation effort is that agreement between humans support the claims of the theory (Teufel, in press). In particular, it may confirm that the phenomena in question indeed exist and that the values attributed to them are clearly defined and distinguishable. The work is mostly of a descriptive nature – it creates phenomenological definitions that encompass bits of observable language.

Our own work is similar to the latter type of annotation in that it is trying to capture a phenomenon that is still under investigation in the linguistic literature. However, it is also different because the categories we use are fixed by language: the quantifiers *some*, *most* and *all* exist and we assume that their definition is agreed upon by speakers of English. What we are trying to investigate is whether those quantifiers should be used at all in the context of ambiguous quantification.

The type of annotation carried out in this paper can be said to have more formal aims than the tasks usually attempted in computational linguistics. In particular, it concerns itself with some of the broad claims made by formal semantics: its model-theoretical view and the use of generalised quantifiers to formalise noun phrases.

In Section 1, we assumed that quantifiers denote relations between sets and presented the task of quantification resolution as choosing the 'correct' set relation for a particular noun phrase in a particular sentence – implying some sort of truth value at work throughout the process: the correct set relation produces the sentence with truth value 1 while the other set relations produce a truth value of 0. What we declined to discuss, though, is the way that those reference sets were selected in natural language, i.e. we didn't make claims about what model, or models, are used by humans when they compute the truth value of a given quantified statement. The annotation task may not answer this question but it should help us ascertain to what extent humans share a model of the world.

In Section 2, we also argued that all subject generic noun phrases could be analysed in terms of quantification. That is, an (underspecified) generalised quantifier is at work in sentences that contain such generic NPs. It is expected that if the annotation is feasible and shows good agreement between annotators, the quantification hypothesis would be confirmed. Thus, annotation may allow us to make semantic claims such as 'genericity does quantify'. Note that the categories we assume are intuitive and do not depend on a particular representation: it is possible to reuse our annotation with a different formalism as long as the theoretical assumption of quantification is agreed upon.

We are not aware of any annotation work in computational linguistics that contributes to validating (or invalidating) a particular formal theory. In that respect, the experiments presented in this paper are of a slightly different nature than the standard research on annotation (despite the fact that, as we will show in the next section, they also aim at producing data for a language analysis system).

### 3.2 Previous work on genericity annotation

The aim of our work being the production of an automatic quantification resolution system, we need an annotated corpus to train and test our machine learning algorithm. There is no corpus that we know of which would give us the required data. The closest contestants are the ACE corpus (2008) and the GNOME corpus (Poesio, 2000) which both focus on the phenomenon of genericity, as described in the linguistic literature. Unfortunately, neither of those corpora are suitable for use in a general quantification task.

The ACE corpus only distinguishes between

'generic' and 'specific' entities. The classification proposed by the authors of the corpus is therefore a lot broader than the one we are attempting here and there is no direct correspondence between their labels and natural language quantifiers: we have shown in Section 2 that genericity didn't map to a particular division of the quantificational space. Furthermore, the ACE guidelines contradict to some extent the literature on genericity. They require for instance that a generic mention be quantifiable with *all*, *most* or *any*. This implies that statements such as *Mosquitoes carry malaria* either refer to a kind only (i.e. they are not quantified) or are not generic at all. Further, despite the above reference to quantification, the authors seem to separate genericity and universal quantification as two antithetical phenomena, as shown by the following quote: "Even if the author may intend to use a GEN reading, if he/she refers to all members of the set rather than the set itself, use the SPC tag".

The GNOME annotation scheme is closer in essence to the literature on genericity and much more detailed than the ACE guidelines. However, the scheme distinguishes only between generic and non-generic entities, as in the ACE corpus case, and the corpus itself is limited to three genres: museum labels, pharmaceutical leaflets, and tutorial dialogues. The guidelines are therefore tailored to the domains under consideration; for instance, bare noun phrases are said to be typically generic. This restricted solution has the advantage of providing good agreement between annotators (Poesio, 2004 reports a Kappa value of 0.82 for this annotation).

## 4 Annotation corpus

We use as corpus a snapshot of the English version of the online encyclopaedia Wikipedia.[1] The choice is motivated by the fact that Wikipedia can be taken as a fairly balanced corpus: although it is presented as an encyclopaedia, it contains a wide variety of text ranging from typical encyclopaedic descriptions to various types of narrative texts (historical reconstructions, film 'spoilers', fiction summaries) to instructional material like rules of games. Further, each article in Wikipedia is written and edited by many contributors, meaning that speaker heterogeneity is high. We would also expect an encyclopaedia to contain relatively many generics, allowing us to assess how our quantificational reading fares in a real annotation task. Finally, the use of an open resource means that the corpus can be freely distributed.[2]

In order to create our annotation corpus, we first isolated the first 100,000 pages in our snapshot and parsed them into a Robust Minimal Recursion Semantics (RMRS) representation (Copestake, 2004) using first the RASP parser (Briscoe et al, 2006) and the RASP to RMRS converter (Ritchie, 2004). We then extracted all constructions of the type Subject-Verb-Object from the obtained corpus and randomly selected 300 of those 'triples' to be annotated. Another 50 random triples were selected for the purpose of annotation training (see Section 7.1).

We show in Figure 1 an example of an annotation instance produced by the parser pipeline. The data provided by the system consists of the triple itself, followed by the argument structure of that triple, including the direct dependents of its constituents, the number and tense information for each constituent, the file from which the triple was extracted and the original sentence in which it appeared. The information provided to annotators is directly extracted from that representation. (Note that the examples were not hand-checked, and some parsing errors may have remained.)

## 5 Evaluating the annotation

In an annotation task, two aspects of agreement are important when trying to prove or refute a particular linguistic model: stability and reproducibility (Krippendorf, 1980). Reproducibility refers to the consistency with which humans apply the scheme guidelines, i.e. to the so-called **inter-annotator agreement**. Stability relates to whether the same annotator will consistently produce the same annotations at different points in time. The measure for stability is called **intra-annotator agreement**. Both measures concern the repeatability of an annotation experiment.

In this work, agreement is calculated for each pair of annotators according to the Kappa measure. There are different versions of Kappa depending on how multiple annotators are treated and how the probabilities of classes are calculated to establish the expected agreement between annotators, $Pr(e)$: we use Fleiss' Kappa (Fleiss, 1971), which allows us to compute agreement between

[1] http://www.wikipedia.org

[2] For access, contact the first author.

```
digraph G211 {
"TRIPLE: weed include pigra" [shape=box];
include -> weed [label="ARG1 n"];
include -> pigra [label="ARG2 n"];
invasive -> weed [label="ARG1 n"];
compound_rel -> pigra [label="ARG1 n"];
compound_rel -> mimosa [label="ARG2 n"];
"DNT INFO: lemma::include() tense::present lpos::v  (arg::ARG1 var::weed() num::pl pos::)
          (arg::ARG2 var::pigra() num::sg pos::)"  [shape=box];
"FILE: /anfs/bigtmp/newr1-50/page101655"  [shape=box];
"ORIGINAL: Invasive weeds include Mimosa  pigra, which covers 80,000 hectares
of the Top End, including vast areas of  Kakadu. " [shape=box]; }
```

Figure 1: Example of annotation instance

multiple annotators.

## 6 An annotation scheme for quantification resolution

### 6.1 Scheme structure

Our complete annotation scheme can be found in Herbelot (to appear). The scheme consists of five parts. The first two present the annotation material and the task itself. Some key definitions are given. The following part describes the various quantification classes to be used in the course of the annotation. Participants are then given detailed instructions for the labelling of various grammatical constructs. Finally, in order to keep the demand on the annotators' cognitive load to a minimum, the last part reiterates the annotation guidelines in the form of diagrammatic decision trees.

In the next sections, we give a walk-through of the guidelines and definitions provided.

### 6.2 Material

Our annotators are first made familiar with the material provided to them. This material consists of 300 entries comprising a single sentence and a triple Subject-Verb-Object which helps the annotator identify which subject noun phrase in the sentence they are requested to label (the 'ORIGINAL' and 'TRIPLE' lines in the parser output – see Figure 1). No other context is provided. This is partly to make the task shorter (letting us annotate more instances) and partly to allow for some limited comparison between human and machine performance (by restricting the amount of information given to our annotators, we force them – to some extent – to use the limited information that would be available to an automatic quantification resolution system, e.g. syntax).

### 6.3 Definitions

In our scheme, we introduce the annotators to the concepts of **quantification** and **kind**.[3]

**Quantification** is described in simple terms, as the process of 'paraphrasing the noun phrase in a particular sentence using an unambiguous term expressing some quantity'. An example is given.

15. *Europeans* discovered the Tuggerah Lakes in 1796 = *Some Europeans* discovered the Tuggerah Lakes in 1796.

We only allow the three quantifiers *some*, *most* and *all*. In order to keep the number of classes to a manageable size, we introduce the additional constraint that the process of quantification must yield a single quantifier. We force the annotator to choose between the three proposed options and introduce priorities in cases of doubt: *most* has priority over *all*, *some* has priority over the other two quantifiers. This ensures we keep a conservative attitude with regard to inference (see Section 1).

**Kinds** are presented as denoting 'the group including all entities described by the noun phrase under consideration', that is, as a supremum. (As mentioned in Section 2, the verbal predicate applies collectively to that supremum in the corresponding formalisation.)

Quantification classes are introduced in a separate part of the scheme. We define the five labels SOME, MOST, ALL, ONE and QUANT (for already quantified noun phrases) and give examples for each one of them.

We try, as much as possible, to keep annotators away from performing complex reference resolution. Their first task is therefore to simply attempt

[3]Distributivity and collectivity are also introduced in the scheme because they are a necessary part of our proposed formalisation. However, as this paper focuses on the annotation of quantification itself, we will not discuss this side of the annotation task.

to paraphrase the existing sentence by appending a relevant quantifier to the noun phrase to be annotated. In some cases, however, this is impossible and no quantifier yields a correct English sentence (this often happens in collective statements). To help our annotators make decisions in those cases, we ask them to distinguish what the noun phrase might refer to when they first hear it and what it refers to at the end of the sentence, i.e., when the verbal predicate has imposed further constraints on the quantification of the NP.

### 6.4 Guidelines

Guidelines are provided for five basic phrase types: quantified noun phrases, proper nouns, plurals, non-bare singulars and bare singulars.

#### 6.4.1 Quantified noun phrases

This is the simplest case: a noun phrase that is already quantified such as *some people*, *6 million inhabitants* or *most of the workers*. The annotator simply marks the noun phrase with a QUANT label.

#### 6.4.2 Proper nouns

Proper nouns are another simple case. But because what annotators understand as a proper noun varies, we provide a definition. We note first that proper nouns are often capitalised. It should however be clear that, while capitalised entities such as *Mary*, *Easter Island* or *Warner Bros* refer to singular, unique objects, others refer to groups or instances of those groups: *The Chicago Bulls*, *a Roman*. The latter can be quantified:

16. The Chicago Bulls won last week. (ALL – collective)
17. A Roman shows courage in battle. (MOST – distributive)

We define proper nouns as noun phrases that 'contain capitalised words and refer to a concept which doesn't have instances'. All proper nouns are annotated as ONE.

#### 6.4.3 Plurals

Plurals must be appropriately quantified and the annotators must also specify whether they are kinds or not. This last decision can simply be made by attempting to paraphrase the sentence with either a definite singular or an indefinite singular – potentially leading to a typical generic statement.

#### 6.4.4 (Non-bare) singulars

Like plurals, singulars must be tested for a kind reading. This is done by attempting to pluralise the noun phrase. If pluralisation is possible, then the kind interpretation is confirmed and quantification is performed. If not (certain non-mass terms have no identifiable parts), the singular refers to a single entity and is annotated as ONE.

#### 6.4.5 Bare singulars

We regard bare singulars as essentially plural, under the linguistic assumption of non-overlapping atomic parts – for instance, water is considered a collection of $H_2O$ molecules, rice is regarded as a collection of grains of rice, etc (see Chierchia, 1998). In order to make this relation clear, we ask annotators to try and paraphrase bare singulars with an (atomic part) plural equivalent and follow, as normal, the decision tree for plurals:

18. *Free software* allows users to co-operate in enhancing and refining the programs they use ≈ *Open source programs* allow users...

When the paraphrase is impossible (as in certain non-mass terms which have no identifiable parts), the noun phrase is deemed a unique entity and labelled ONE.

## 7 Implementation and results

### 7.1 Task implementation

Three annotators were used in our experiment. One annotator was one of the authors; the other two annotators were graduate students (non-linguists), both fluent in English. The two graduate students were provided with individual training sessions where they first read the annotation guidelines, had the opportunity to ask for clarifications, and subsequently annotated, with the help of the author, the 50 noun phrases in the training set. The actual annotation task was performed without communication with the scheme author or the other annotators.

### 7.2 Kappa evaluation

We made an independence assumption between quantification value and kind value, and evaluated agreement separately for each type of annotation.

Intra-annotator agreement was calculated over the set of annotations produced by one of the authors. The original annotation experiment was reproduced at three months' interval and Kappa was

| Class | Kind | Quantification |
|---|---|---|
| Kappa | 0.85 | 0.84 |

Table 1: Intra-annotator agreements for both tasks

| Class | Kind | Quantification |
|---|---|---|
| Kappa | 0.67 | 0.72 |

Table 2: Inter-annotator agreements for both tasks

computed between the original set and the new set. Table 1 shows results over 0.8 for both tasks, corresponding to 'perfect agreement' according to the Landis and Koch classification (1977). This indicates that the stability of the scheme is high.

Table 2 shows inter-annotator agreements of over 0.6 for both tasks, which correspond to 'substantial agreement'. This result must be taken with caution, though. Although it shows good agreement overall, it is important to ascertain in what measure it holds for separate classes. In an effort to report such per class agreement, we calculate Kappa values for each label by evaluating each class against all others collapsed together (as suggested by Krippendorf, 1980).

Table 3 indicates that substantial agreement is maintained for separate classes in the kind annotation task. Table 4, however, suggests that, if agreement is perfect for the ONE and QUANT classes, it is very much lower for the SOME, MOST and ALL classes. While it is clear that the latter three are the most complex to analyse, we can show that the lower results attached to them are partly due to issues related to Kappa as a measure of agreement. Feinstein and Cicchetti (1990), followed by Di Eugenio and Glass (2004) proved that Kappa is subject to the effect of prevalence and that different marginal distributions can lead to very different Kappa values for the same observed agreement. It can be shown, in particular, that an unbalanced, symmetrical distribution of the data produces much lower figures than balanced or unbalanced, asymmetrical distributions because the expected agreement gets inflated. Our confusion matrices indicate that our data falls into the category of unbalanced, symmetrical distribution: the classes are not evenly distributed but annotators agree on the relative prevalence of each class. Moreover, in the quantification task itself, the ONE class covers roughly 50% of the data. This means that, when calculating per class agreement, we get an approximately balanced distribution for the ONE label and an unbalanced, but still symmetrical, distribution for the other labels. This leads to the expected agreement being rather low for the ONE class and very high for the other classes. Table 5 reproduces the per class agreement figures obtained for the quantification task but shows, in addition, the observed and expected agreements for each label. Although the observed agreement is consistently close to, or over, 0.9, the Kappa values differ widely in conjunction with expected agreement. This results in relatively low results for SOME, MOST and ALL (the QUANT label has nearly perfect agreement and therefore doesn't suffer from prevalence).

| Class | KIND | NOT-KIND | QUANT |
|---|---|---|---|
| Kappa | 0.63 | 0.71 | 0.88 |

Table 3: Per class inter-annotator agreement for the kind annotation

| Class | ONE | SOME | MOST | ALL | QUANT |
|---|---|---|---|---|---|
| Kappa | 0.81 | 0.45 | 0.44 | 0.51 | 0.88 |

Table 4: Per class inter-annotator agreement for the quantification annotation

| Class | Kappa | Pr(a) | Pr(e) |
|---|---|---|---|
| ONE | 0.814 | 0.911 | 0.521 |
| SOME | 0.445 | 0.893 | 0.808 |
| MOST | 0.438 | 0.931 | 0.877 |
| ALL | 0.509 | 0.867 | 0.728 |
| QUANT | 0.884 | 0.987 | 0.885 |

Table 5: The effect of prevalence on per class agreement, quantification task. $Pr(a)$ is the observed agreement between annotators, $Pr(e)$ the expected agreement.

With regard to the purpose of creating a gold standard for a quantification resolution system, we also note that out of 300 quantification annotations, there are only 14 cases in which a majority decision cannot be found, i.e., at least two annotators agreed in 95% of cases. Thus, despite some low Kappa results, the data can adequately be used for the production of training material.[4]

[4] As far as such data ever can be: Reidsma and Carletta, 2008, show that systematic disagreements between annotators will produce bad machine learning, regardless of the Kappa obtained on the data.

In Section 8, we introduce difficulties encountered by our subjects, as related in post-annotation discussions. We focus on quantification.

## 8 Annotation issues

### 8.1 Reference

Although we tried to make the task as simple as possible for the annotators by asking them to paraphrase the sentences that they were reading, they were not free from having to work out the referent of the NP (consciously or unconsciously) and we have evidence that they did not always pick the same referent, leading to disagreements at the quantification stage. Consider the following:

19. Subsequent annexations by Florence in the area have further diminished the likelihood of incorporation.

In the course of post-annotation discussions, it became clear that not all annotators had chosen the same referent when quantifying the subject NP in the first clause. One annotator had chosen as referent *subsequent annexations*, leading to the reading *Some subsequent annexations, conducted by Florence in the area, have further diminished the likelihood of incorporation*. The other two annotators had kept the whole NP as referent, leading to the reading *All the subsequent annexations conducted by Florence in the area have further diminished the likelihood of incorporation*.

### 8.2 World knowledge

Being given only one sentence as context for the NP to quantify, annotators sometimes lacked the world knowledge necessary to make an informed decision. This is illustrated by the following:

20. The undergraduate schools maintain a non-restrictive Early Action admissions programme.

Discussion revealed that all three annotators had a different interpretation of what the mentioned Early Action programme might refer to, and of the duties of the undergraduate schools with regard to it. This led to three different quantifications: SOME, MOST and ALL.

### 8.3 Interaction with time

The existence of interactions between NP quantification and what we will call temporal quantification is not surprising: we refer to the literature on genericity and in particular to Krifka et al (1995) who talk of characteristic predication, or habituality, as a phenomenon encompassed by genericity. We do not intend to argue for a unified theory of quantification, as temporal quantification involves complexities which are beyond the scope of this work. However, the interactions observed between temporality and NP quantification might explain further disagreements in the annotation task. The following is a sentence that contains a temporal adverb (*sometimes*) and that produced some disagreement amongst annotators:

21. Scottish fiddlers emulating 18th-century playing styles sometimes use a replica of the type of bow used in that period.

Two annotators labelled the subject of that sentence as MOST, while the third one preferred SOME. In order to understand the issue, consider the following, related, statement:

22. Mosquitoes sometimes carry malaria.

This sentence has the possible readings: *Some mosquitoes carry malaria* or *Mosquitoes, from time to time in their lives, carry malaria.* The first reading is clearly the preferred one.

The structure of (21) is identical to that of (22) and it should therefore be taken as similarly ambiguous: it either means that some of the Scottish fiddlers emulating 18th-century playing styles use a replica of the bow used in that period, or that a Scottish fiddler who emulates 18th-century playing styles, from time to time, uses a replica of such a bow. The two readings may explain the labels given to that sentence by the annotators.

## 9 Conclusion

Taking prevalence effects into account, we believe that our agreement results can be taken as evidence that underquantification is analysable in a consistent way by humans. We also consider them as strong support for our claim that 'genericity quantifies'. Our scheme could however be refined further. In a future version, we would add guidelines regarding the selection of the referent of the noun phrase, encourage the use of external resources to obtain the context of a given sentence (or simply provide the actual context of the sentence), and give some pointers as to how to resolve issues or ambiguities caused by temporal quantification.

## References


ACE. 2008. *ACE (Automatic Content Extraction) English Annotation Guidelines for Entities*, Version 6.6 2008.06.13. Linguistic Data Consortium.

Edward Briscoe, John Carroll and Rebecca Watson. 2006. 'The Second Release of the RASP System'. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, Sydney, Australia, 2006.

Gregory Carlson. 1995. 'Truth-conditions of Generics Sentences: Two Contrasting Views'. In Gregory N. Carlson and Francis Jeffrey Pelletier, Editors, *The Generic Book*, pages 224 – 237. Chicago University Press.

Gennaro Chierchia. 1998. 'Reference to kinds across languages'. *Natural Language Semantics*, 6:339–405.

Ariel Cohen. 1996. *Think Generic: The Meaning and Use of Generic Sentences*. Ph.D. Dissertation. Carnegie Mellon University at Pittsburgh. Published by CSLI, Stanford, 1999.

Ann Copestake. 2004. 'Robust Minimal Recursion Semantics'. `www.cl.cam.ac.uk/~aac10/papers/rmrsdraft.pdf`.

Barbara Di Eugenio and Michael Glass. 2004. 'The kappa statistic: a second look'. *Computational Linguistics*, 30(1):95–101.

Alvan R. Feinstein and Domenic V. Cicchetti. 1990. 'High agreement but low kappa: I. The problems of two paradoxes'. *Journal of Clinical Epidemiology*, 43(6):543–549.

Joseph Fleiss. 1971. 'Measuring nominal scale agreement among many raters'. *Psychological Bulletin*, 76(5):378-382.

Aurelie Herbelot. To appear. *Underspecified quantification*. Ph.D. Dissertation. Computer Laboratory, University of Cambridge, United Kingdom.

Gerhard Heyer. 1990. 'Semantics and Knowledge Representation in the Analysis of Generic Descriptions'. *Journal of Semantics*, 7(1):93–110.

Paul Kingsbury, Martha Palmer and Mitch Marcus. 2002. 'Adding Semantic Annotation to the Penn TreeBank'. In *Proceedings of the Human Language Technology Conference (HLT 2002)*, San Diego, California, pages 252–256.

Manfred Krifka, Francis Jeffry Pelletier, Gregory N. Carlson, Alice ter Meulen, Godehard Link and Gennaro Chierchia. 1995. 'Genericity: An Introduction'. In Gregory N. Carlson and Francis Jeffry Pelletier, Editors. *The Generic Book*, pages 1–125. Chicago: Chicago University Press.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Newbury Park, CA: Sage.

J. Richard Landis and Gary G. Koch. 1977. 'The Measurement of Observer Agreement for Categorical Data'. *Biometrics*, 33:159–174.

Sara-Jane Leslie. 2008. 'Generics: Cognition and Acquisition.' *Philosophical Review*, 117(1):1–47.

Christopher Lyons. 1999. *Definiteness*. Cambridge University Press, Cambridge, UK.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. 'Building a large annotated corpus of english: The penn treebank'. *Computational Linguistics*, 19(2):313–330.

Francis Jeffry Pelletier and Nicolas Asher. 1997. 'Generics and defaults'. In: Johan van Benthem and Alice ter Meulen, Editors, *Handbook of Logic and Language*, pages 1125–1177. Amsterdam: Elsevier.

Massimo Poesio. 2000. 'The GNOME annotation scheme manual', Fourth Version. `http://cswww.essex.ac.uk/Research/nle/corpora/GNOME/anno_manual_4.htm`

Massimo Poesio. 2004. 'Discourse Annotation and Semantic Annotation in the GNOME Corpus'. In: *Proceedings of the ACL Workshop on Discourse Annotation*, Barcelona, Spain.

Dennis Reidsma and Jean Carletta. 2008. 'Reliability measurement without limits'. *Computational Linguistics*, 34(3), pages 319–326.

Anna Ritchie. 2004. 'Compatible RMRS Representations from RASP and the ERG'. `http://www.cl.cam.ac.uk/TechReports/UCAM-CL-TR-661`.

Simone Teufel, Advaith Siddharthan, Dan Tidhar. 2006. 'An annotation scheme for citation function'. In: *Proceedings of Sigdial-06*, Sydney, Australia, pages 80–87.

Simone Teufel. 2010. *The Structure of Scientific Articles: Applications to Summarisation and Citation Indexing*. CSLI Publications. In press.